Journal of the Maharaja Sayajirao University of Baroda
ISSN :0025-0422

A NOVEL ENSEMBLE CLASSIFICATION TECHNIQUE METHOD TO IMPROVE THE PREDICTION OF CARDIOVASCULAR DISEASE

C. Keerthana, Assistant Professor, Department of computer science, Nallamuthu Gounder Mahalingam College, Pollachi, Tamilnadu - kirthy6@gmail.com
Dr.B.Azhagusundari, Associate professor, Department of computer science, Nallamuthu Gounder Mahalingam College, Pollachi, Tamilnadu - azhagumithra78@gmail.com

Abstract: Machine learning is a subset of artificial intelligence used to handle a variety of data science problems. In machine learning applications, the prediction of an outcome based on existing data is common. The system learns patterns from an existing dataset and then applies them to an unknown dataset to predict the outcome. Some classification techniques forecast reasonably well, while others have limits. This study investigates ensemble classification, a technique that combines many classifiers to increase the accuracy of weak algorithms. Experiments were carried out on a database of people with cardiac disease. The ensemble technique is employed to increase predicting accuracy in heart disease using a relative analytical approach in this paper. This research focuses not only on improving the accuracy of weak classification algorithms, but also on demonstrating the algorithm's utility in early disease prediction using a medical dataset. The use of ensemble classification resulted in a maximum increase in accuracy of 9% for weak classifiers. The approach was improved further by incorporating feature selection, which resulted in a significant increase in prediction accuracy.
Keywords: Heart disease ,Machine learning, Ensemble methods, classification, Prediction model

INTRODUCTION:

Heart disease is one of the most common diseases afflicting many people in their middle or later years, and it frequently results in fatal complications [1]. According to WHO data, heart problems account for 24% of non-communicable disease mortality in India [2]. One-third of all deaths globally are caused by heart disease [2,3]. Every year, over 19 million individuals die as a result of cardiovascular disease (CVD) [4, 5]. The Cleveland Heart Disease Resource (CHDD) is widely regarded as the most comprehensive database for research on heart disease [5]. Coronary heart disease, angina pectoris, congestive heart failure, cardiomyopathy, congenital heart disease, arrhythmias, and myocarditis are all examples of cardiovascular diseases. It is difficult to forecast the likelihood of heart disease based just on risk factors [5].To forecast the outcome of existing data, a machine learning technique can be used. This paper predicts heart disease risk using a machine learning technique known as classification based on risk factors. An ensemble technique is also utilised to increase the accuracy of forecasting heart disease risk.

LITERATURE SURVEY

Machine learning and data mining are widely used in a range of fields. Due of the vastness of healthcare data resources, managing them manually is difficult [1]. Some of the techniques used for such prediction issues include Support Vector Machines (SVM), Neural Networks, Decision Trees, Regression, and Naive Bayes classifiers. According to [6], the best predictor was SVM, which had 92.1% accuracy, followed by neural networks (91%) and decision trees (89.6%). SVM, neural networks, decision trees, Naive Bayes, and associative categorization are all highly effective in predicting heart disease. Associative categorization outperforms unstructured data in terms of accuracy and flexibility [4, 7].A review of classification approaches revealed that Naive Bayes was the best algorithm, followed by neural networks and decision trees [5]. Back propagation algorithms were employed to train supervised networks to identify heart illness, and the findings were accurate [8]. Feature extraction utilizing evolutionary algorithms and neural networks based on fuzzy logic was more accurate than 94.79% [9]. Set-based classification with a heterogeneous dataset produced classification precision of up to 93.5% [10]. Heartbeat classification using SVM-based classifiers has

proven to be exceptionally accurate, and overall performance of the classifier has been improved via particle swarm optimization [3, 11].

METHODOLOGY

DATASET DESCRIPTION

We used a Cleveland dataset with the 13 parameters provided in Table 1 for this analysis. We refined the data after consolidating it into a single database by finding distorted data, such as contradicting entries or missing values [12, 13]. We omitted some data points at this moment in order to make more accurate estimates. After that, the revised data was divided into two groups: testing and training. Booting, Bagging, and Stacking ensemble strategies were utilized to implement the selected data in the training set.

Table 1: Cleveland Dataset

Features	Description	Features	Description
Age	Age (in years)	Exang	exercise induced angina
Sex	Gender	Oldpeak	ST depression induced by exercise relative to rest
Cp	Chest pain type	Slope	the slope of the peak exercise ST segment
thalach	maximum heart rate achieved	Ca	number of major vessels (0-3)
Chol	serum cholesterol in mg/dl	Thal	3 = normal; 6 = fixed defect; 7 = reversable defect
Fbs	Fasting blood sugar	Restecg	resting electrocardiographic results

ALGORITHMS AND CLASSIFICATION

Classification is a supervised learning method that predicts outcomes based on previously acquired data [13]. The researchers proposed using classification algorithms to diagnose cardiac disease and employing an ensemble of classifiers to improve classification accuracy. Individual classifiers were trained using the Cleveland dataset, which is broken into two sections: training and test dataset. The test dataset was used to assess the classifier's performance.

RANDOM FOREST

A random forest is a tree-based classification method. This strategy generates a forest with a large number of trees. It is a hybrid algorithm that combines several algorithms. A sampling approach of the training dataset is used to build a collection of decision trees. It repeats the operation with multiple random samples until a predicted ensemble with majority voting is reached. Though the random forest method was effective in coping with missing variables, obtaining a precise number was difficult. Appropriate parameter adjustment may be utilised to avoid overfitting [12]. The Random Forest Algorithm is depicted in Figure 1.

Input:
Training Dataset D
Set of s original features
$G=\{g_1,g_2 \ldots g_s\}$
Output:
Feature subset
Code:
Final ranking F
Repeat for j in {1:s-1},
Rank set R using random forest
$g^* \leftarrow$ last ranked feature in R
F(s-j+1) g
**$R \leftarrow R - g^*$*

Figure 1. Algorithm for Random Forest.

Let E be a node
Let T be a tuple in class A and Y be the set of attributes $Z=\{1, 2,\ldots, n\}$
If $T \subset A$
return $E=L(A)$,*where L is a leaf node, if* $Z = \phi$
then
return E as a leaf node with the class of majority in T , E is divided with the best splitting criterion for each splitting criterion i
Ti –set of tuples satisfying I, if $Ti= \phi$ *then*
combine a leaf node in T class to E node;
else
combine the node return by decision tree(Ti, attribute list)to E node;
end for
return E

Figure 2. Algorithm for C4.5

The C4.5 algorithm is based on the Iterative Dichotomiser 3 (ID3) algorithm, which is a classification tree-based technique. This algorithm was created by Quinlan. It divides the trees based on the information gain ratio. It accepts data as input and produces a decision tree as output. This method is used to generate univariate trees. Classification rules are represented using decision trees. When a tree's split falls below a certain threshold value, it is stopped. It employs the pruning method to reduce inaccuracy and is a great tool for dealing with numerical properties [14]. As shown in Fig. 2, the C4.5 algorithm is utilised to generate a decision tree from training tuples.

MULTILAYER PERCEPTRON

Artificial neurons having several layers, including hidden layers, were employed in the Multilayer Perceptron (MLP) method [15]. This algorithm was used to solve a number of binary classification problems. A perceptron's neurons each have an activation function. Multilayer perceptron evolved from genetic neurons as computational architectures. By mapping each neuron's weighted inputs, the activation function reduces the number of layers to two. A perceptron learns by changing the weights allocated to it. Equ (1) and (2) determined the outputs of each neuron in the hidden layer, which are as follows:

$$o(x) = G\, b(2) + W(2)\, h(x)) \quad (1)$$

$$h(x) = \phi(x) = H\,(b(1) + W(1)\, x) \quad (2)$$

where b(1), b(2) are bias vectors; W(1), W(2) are weight matrices and activation functions G and H. The set of parameters to learn is the set $\theta = \{W(1), b(1), W(2), b(2)\}$.

NAIVE BAYES

Naive Bayes is a mathematical categorization approach based on Bayes' theorem. It is assumed that each characteristic and variable has a varied prognosis and prevalence [6]. Each characteristic in the test data and objective was computed using the prior probability of Bayes theory, and the target with the highest probability was chosen as a consequence [2]. The probability can be calculated using (3)

$$P(Cj|Fi) = P(Fi|Cj)P(Cj) / P(Fi) \quad (3)$$

Where P(Cj|Fi) is the probability of a specific class, (Cj) appearing with a explicit feature (Fi) from the total of all Features F and Classes C. P(Cj) is the prospect of a definite class (Cj) appearing with a specific feature (Fi) from all classes (C), P(Fi |Cj) is the probability of a specific feature (Fi) appearing with a specific class (Cj) from sum of all features (F).

BAYES NET

The Bayesian network is one of the probability-based prediction models that uses a graphical manner. Using discrete and continuous information, it forecasts and diagnoses problems. A set of variables with conditional relationships defined as acyclic directed graphs characterizes this net. Edges between nodes in a Bayes net indicate subordinate qualities, but nodes that are not related are conditionally in-dependent. Assume A represents a fact with n attributes (A = A1, A2,....., An) and G represents the hypothesis that facts belong to the B class. P(G|A) is the probability of hypothesis G given the facts A. P(A|G) is the subsequent probability of A trained on G. As shown in Equation, the Bayes net can be estimated using possibility (4).

$$P(G|A) = P(A|G)P(G)/P(A) \quad (4)$$

P(G) is the probability that the hypothesis is correct and P(A) is the probability facts. P(A|G) is the probability fact that hypothesis given is correct, and P(G|A) the possibility of the hypothesis if the evidence is correct.

SVM

Support Vector Machines have demonstrated outstanding performance in disease prediction in recent years [4]. SVM is a supervised learning technique that aims to reduce generalization errors when conducting regression and classification tasks. SVM is very effective in high-dimensional domains, and it is scientifically represented by Equ (5), (6), and (7).

$$\text{If } Yi = +1; wxi + b \geq 1 \quad (5)$$

$$\text{If } Yi = -1; wxi + b \leq 1 \quad (6)$$

$$\text{For all i; } yi(wi + b) \geq 1 \quad (7)$$

Where x is a point of vector in a hyper plane and w is a weight of each vector. To discriminate the data in Equ (5) &(6), the data in Equ (5) must be superior to zero and in Equ (6) the data must be less than zero. SVM chooses the hyperplane with the largest distance out of all the possibilities.

ENSEMBLE METHODS

The ensemble method is used to increase classification accuracy. This method employs a data inside data categorization methodology to connect fragile learners with sturdy learners in order to boost the fragile learner's efficiency. In this research, various ensemble approaches are employed to improve the accuracy of heart disease prediction by combining many classifiers to reach higher accuracy than individual classifiers.

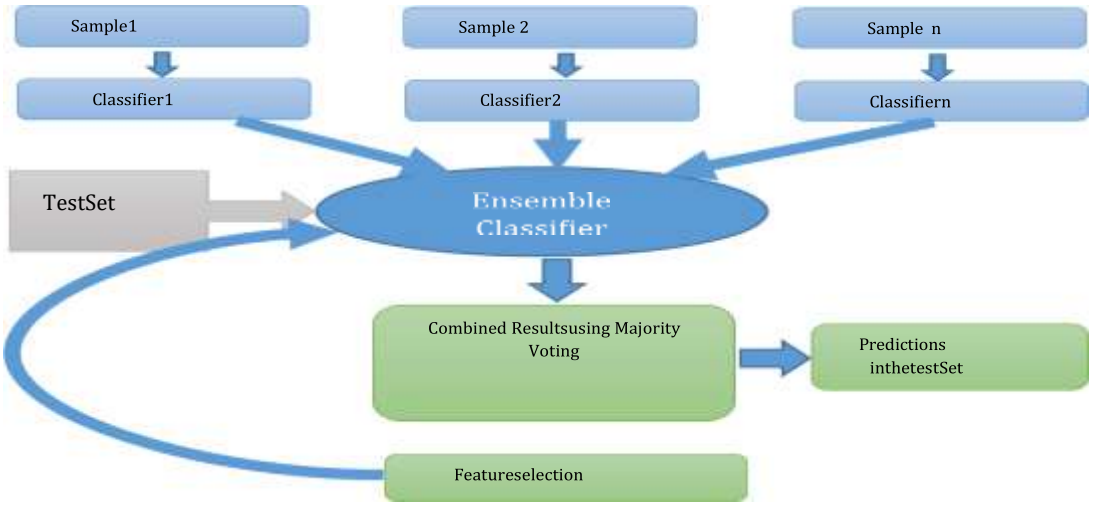

Figure 3 depicts an ensemble procedure

BOOSTING

Boosting is one of the ensemble meta-algorithm techniques used to eliminate bias. In order to enhance speed, the data is partitioned into multiple subclasses when boosting. The subclass is used to train the classifier, resulting in a series of models with low performance. The objects that the previous model could not accurately classify are used to create new subsets. The assembling procedure then enhances their performance by merging the weak models by employing a cost function. Figure 4 displays the Boosting Algorithm.

Let D={d_1,d_2,d_3,…d_n} be the given dataset E={}, the set of ensemble classifiers
C={c_1,c_2,c_3,…c_n},the set of classifiers X=the training set ,XD
Y = the test set, YDL=n(D)
Let init=1
S(init)=A random subset of X;S(init)XM(0)={}
For i=1toLdo if i>1
s(i)=Set of incorrectly classified instances of M(i-1)+S(i)M(i)=Model trained using C(i) on S(i)
E=EC(i)
end if next i
for i=1toL
R(i)=Y classified by E(i)next i
Result=max(R(i):i=1,2,…,n)

Figure 4. Boosting Algorithm

BAGGING

Bagging is also known as bootstrap aggregation. Bagging replaces a pattern from the training set at random. The new training set contains the same amount of patterns as the old training set. The new training set's name is Bootstrap replicate. Bagging entails obtaining bootstrap samples and training each sample with various classifiers. The votes of all classifiers were combined up, and the outcome of each classification was determined by the majority vote method, which was average. According to research, bagging is used to improve the act of a bad learner to its full capacity. Figure 5 depicts the Bagging Algorithm.

Let D = {d_1, d_2, d_3, … d_n} be the given dataset E={},the set of ensemble classifiers
C={c_1,c_2,c_3,…c_n},the set of classifiers X = the training set, X
DY = the test set, Y DL=n(D)
for i=1to L do
S(i) = {Bootstrap sample I with replacement}I
XM(i)=Model trained using C(i) on S(i)
E=EC(i)
next i
for i=1toL
R(i)=Y classified by E(i)next i
Result=max(R(i):i=1,2,…,n)

Figure 5. Bagging Algorithm

STACKING

Stacking is a type of Meta classifier that mixes many classification models. Several layers are piled on top of one another. Each model sends its predictions to the one above, and the top layer takes decisions

based on the models below. The base layer models use input features from the original dataset. The highest layer makes the forecast, which gets its information from the base layer. The Stacking Algorithm is depicted in Figure 6. In stacking, the unique data is used to load a variety of different patterns. The patterns with the greatest effects were picked, while the rest were rejected. Stacking combines multiple base classifiers learned using different learning techniques M on a single dataset D using a Meta classifier.

MAJORITY VOTING CLASSIFIER

A majority classifier is a meta-classification that uses a majority vote method to group each classifier together. The group tag has various classifier's majority vote to predict the ending group tag. The final class label fj is defined as

$$fj = \text{mode } \{C1, C2, \ldots, Cn\}$$

Where {C1, C2, …, Cn} denotes the individual classifiers involved in the voting. Figure 7 depicts the Majority Voting Algorithm.

Let $D=\{d_1,d_2,d_3,\ldots d_n\}$be the given dataset
$E=\{E_1,E_2,E_3,\ldots E_n\}$,the set of ensemble classifiers $C=\{c_1,c_2,c_3,\ldots c_n\}$, the set of classifiers
Y=the training set,
YDZ=the test set, Z
K=meta level classifier
L=n(D)
for j=1toLdo
M(j)=Model trained using
E(j)on XN ext j
M=MK
Result=Z classified by M

Figure 6. Stacking Algorithm

Let be the prediction of the ith classifier on a class with j labels

$$\sum_{i=1}^{n} c_{ij} = \max_{j=1,\ldots,m} \sum_{i=1}^{n} c_{ij}$$

The ensemble classifier's probability for the decision to be better is

$$P_{ens} \sum_{k=\frac{0}{2}+1}^{n} \binom{0}{k} p^{k}(1-p)^{n-k}$$

Figure 7. Majority Voting Algorithm

EXPERIMENTALRESULTS

ENSEMBLE PERFORMANCE WITH CLASSIFIER'S

An examination of various classification techniques was performed using the Cleveland dataset. Some algorithms work well, while others fail miserably. Ensemble approaches are utilised in this paper to improve the performance of fragile classifiers. In this work, many ensemble algorithms are applied. Ensembles are built using machine learning methods such as Support Vector Machine, Naive Bayes, Random Forest, Bayesian network, C4.5, and Multilayer Perceptron. Classifiers rely on majority voting to determine detection accuracy. The Naive Bayes classifier functions as a stacking meta-classification approach, with results collected by stacking with three or four additional classifiers, accordingly.

The outcome suggest when fragile classifiers are ensembled, they execute superior than the previous results. The classification of the dataset is done with the R tool.

During the preprocessing stage, the Cleveland dataset is cleaned and checked for duplicate values, missing and baseless data. With this dataset, many classifier techniques are utilised. Bayesian Network, C4.5, and Multilayer Perceptron were shown to be poorer than Naive Bayes, SVM, and Random Forest. Although ensemble classifiers are well-known for boosting classification accuracy, meta-classification methods have been used to test poor learners. The ensemble approach was used to assess

the results of its various techniques, including bagging, boosting, stacking, and majority voting. The performance of the classification model was assessed using ten-fold cross validation. The complete dataset is partitioned into 10 subsets and processed ten times in this method, with nine subsets allocated as training models and enduring subset as test model. The outcome was calculated with aggregating the findings from ten iterations.

In Figure8, Individual base classifier is compared with bagging algorithm. As applying individual classifiers to classify the dataset, the accuracy rates of Naive Bayes, Random forest, Bayes Net, C4.5, Multilevel Perceptron, and SVM range from 73.58 % to 85.07%. The SVM classifier has the highest accuracy of 85.07%, while Bayes Net, Multilevel Perceptron, and C4.5 all have lower than 78% accuracy. The results indicate that using bagging algorithm improves its classification precision with 7.87%.

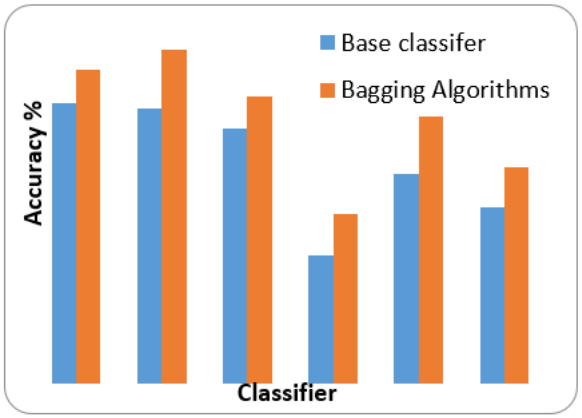

Figure 8. Bagging Accuracy

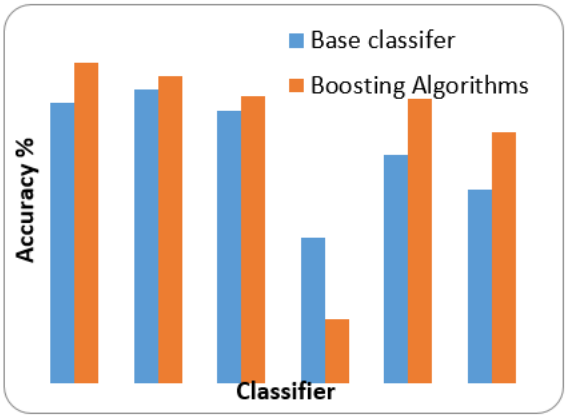

Figure 9. Boosting Accuracy

The outcomes of the ensemble technique, using boosting, are shown in Figure 9. Using boosting, the Naive Bayes algorithm obtained 0.99 %, the Bayes Net obtained 2.65%, the multilayer perceptron obtained 2.87%, and SVM gained 1.04%. With boosting, the Random Forest algorithm is raised with highest value.

According to Figure 10, combining fragile classifiers with sturdy classifiers and applying popular voting enhances the correctness of fragile classifier significantly. When combined with strong classifiers, the C4.5 algorithm can improve accuracy by 7.26%. The accuracy of Bayes Net was improved by 4.66% after it was ensembled with sturdy classifier subsets. The exactness of multilayer perceptron with sturdy classifier was enhanced with 1.25%.

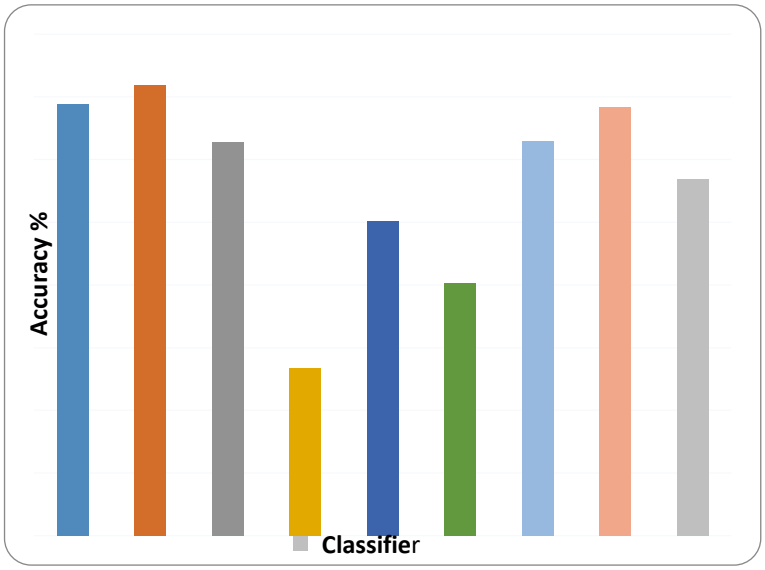

Figure 10. Classifier with Majority Voting

Stacking is an ensemble method that involves stacking with the lowest level of classifiers and a meta-classification. In this work, random tree and random forest classifiers were used with Meta classification. SVM, Naive Bayes, Bayes Net, and C4.5 are the fundamental classifiers. As shown in Figure 11, the random forest classifier outperforms the random tree classifier in terms of precision. During stacking technique C4.5 classifier improves their accuracy by 1.87 %, while all other algorithms decreased accuracy by up to 2.45%. The accurateness of fragile classifiers was enhanced, when they were stacking with random forest classifier. The Bayesian network's correctness samplified with 0.79 %, C4.5 with 2.18%, the Multilayer Perceptron with 3.02%, and SVM with 6.29%. The result implies the accuracy of fragile classifiers were higher when they are stacked with random forest classifier.

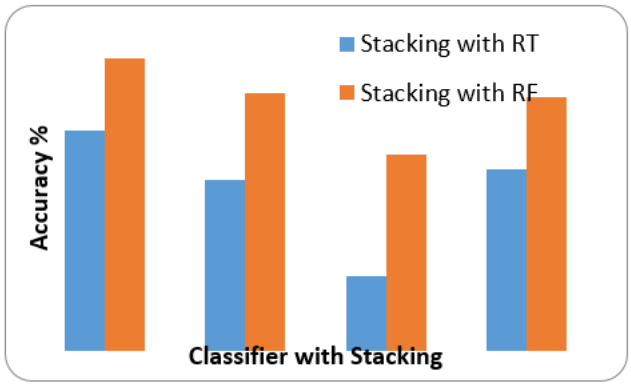

Figure 11. Stacking with RF &RT

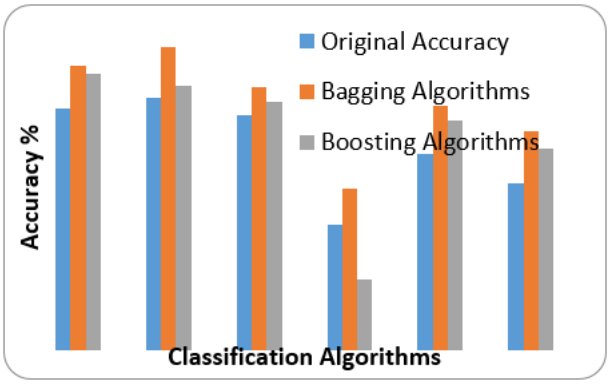

Figure 12. Comparing Bagging with Boosting

Figure 12 depicts the contrast bagging and boosting procedure. According to the findings, both bagging and boosting algorithms are effective at improving the precision of bad classifiers. Bagging helps all fragile classifiers function better.

The exactness of fragile classifiers is increased by up to 6.88% when the various ensemble approaches are compared. Figure 13 shows the largest increase in precision of a fragile classifier using several ensemble approaches. The results indicate that an ensemble strategy is superior in improving the exactness of fragile classifiers, with voting producing superior outcomes.

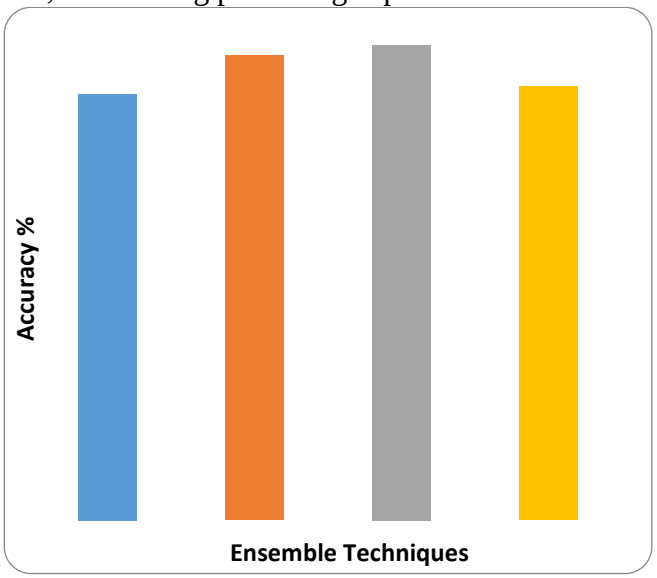

Figure 13. Comparison of Ensemble Methods

Using feature selection, the classifier's accuracy is enhanced even further [4]. For the purposes of evaluating the performances, six sets of features were chosen. The qualities 'age' and 'sex' are regarded personal information, while the remaining 11 traits are gathered through the patient's medical observation. Using Cleveland dataset, mixture of three attributes were selected among the attributes in the dataset. Individual combination is assessed with various classifiers. Then testing was continual to discover best four-attribute grouping out of thirteen in total. Without taking the empty set into account,

the maximum number of combinations from 13 qualities is 2n -1.The combination of less than three qualities is ignored in this experiment. As a result, the overall combinations was calculated using,

$$2^n - \left[\frac{n^2 + n}{2} + 1 \right]$$

Where n is a count of overall attributes. Combinations of feature sets are named as FS1, FS2, FS3, FS4, FS5 and FS6. Following is a list of the features and their descriptions:
FS1 = {sex, ca, thalach, exang, fbs, slope, cp}
FS2 = {ca, thalach, cp, exang, oldpeak, sex}
FS3 = {ca, oldpeak, sex, cp, restecg, fbs, thal}
FS4 = {age, sex, chol, ca, fbs, oldpeak, slope,exang, cp}
FS5 = {restecg, sex, ca, chol, age, oldpeak, slope, cp, thal}
FS6 = {sex, ca, trestbps, fbs, oldpeak, thalach, exang, restecg, slope, cp, thal}
Table 2 shows the improvement in bagging accuracy as a result of feature selection.

Table 2. Improvement in bagging accuracy

Enhancement in Bagging Method with Feature Set Selection			
Algorithm used	Accuracy in Bagging	Enhance accuracy with Feature Set Selection	Feature Set
Bayes Net	83.66	83.92	FS3
Random Forest	79.42	80.18	FS4
Random Forest	79.42	83.62	FS6
C4.5	78.35	83.95	FS3
Multilayer Perceptron	81.29	81.28	FS2
Multilayer Perceptron	81.29	83.78	FS6
Multilayer Perceptron	81.29	82.25	FS1
Naïve Bayes	83.06	83.72	FS6

Using bagging, C4.5 classifier with feature selection set FS3, had largest gain with precision of 2.31%. Feature selection sets of FS2 and FS6 improved the multilayer perceptron exactness by 0.66%. With feature selection set FS6, the accuracy of random forest improved with 2.56%.
According to the result, the exactness of boosting algorithm was enhanced with greatest level of 3.11% throughC4.5 classifier and feature selection set FS6. With feature selection set FS6, the highest raise in boosting exactness with random forest was 3.3%. With feature set FS2, there was a 1.32% increase in boosting with multilayer perceptron. With feature set FS6, there was a 0.33% increase in the Naive Bayes classifier with boosting. As a result, feature selection set FS6 shows improvement in forecast of different classifiers with Random forest, C4.5, and Naive Bayes. Table 3 summarizes the result.

Table 3. Result

Enhancement in Boosting Method with Feature Set Selection			
Algorithm used	Bagging Accuracy	Increase in accuracy with Feature Selection	Feature Set
C4.5	75.9	79.87	FS6
C4.5	75.9	79.21	FS1
C4.5	75.9	78.22	FS4
C4.5	75.9	77.23	FS5
C4.5	75.9	76.57	FS2
Random Forest	78.88	82.18	FS6
Random Forest	78.88	80.86	FS4
Random Forest	78.88	80.86	FS1

Random Forest	78.88	79.87	FS5
Multilayer Perceptron	79.54	80.86	FS2
Multilayer Perceptron	79.54	80.53	FS5
Naïve Bayes	84.16	84.49	FS6

Feature selection improves majority voting as well. The entire feature selection sets are enhanced with precision of Bayesian Network, Majority Voting, Naive Bayes, Random Forest, and Multilayer Perceptron. The greatest improvement of 3.53% with feature selection set FS4 was acquired. Figure 14 depicts improvement in precision of ensemble majority voter with SVM, Naive Bayes, Random Forest, and Multilayer Perceptron classifiers. Feature set FS4 provided the greatest increase in accuracy. Figure 15 depicts the increase in majority voting accuracy of SVM, Naive Bayes, Random Forest, and C4.5.Feature selection set FS6 caused greatest increase with accuracy.

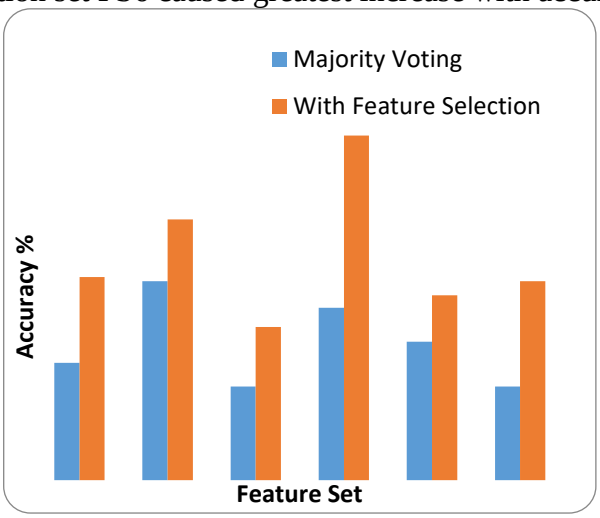

Figure 14. Improvement in precision of Majority Voting with RF, C4.5 and RF using Feature set Selection

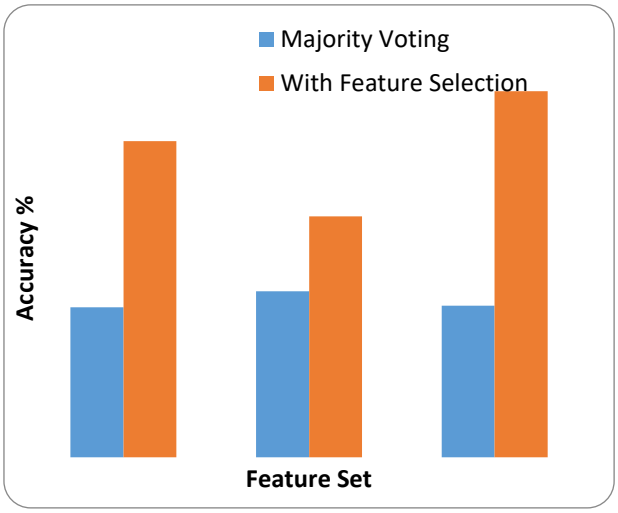

Figure 15. Improvement in precision of Majority Voting with SVM, NB, SVM, NB and MLP using Feature Set Selection

Figure 16 depicts improvement with precision of ensemble majority voter method using Naive Bayes, SVM, Bayesian Network and Random Forest. Feature selection shows enhancement in stacking as well. Stacking Naive Bayes, Bayesian Network, C4.5, and MLP with Random Forest improves 0.96% with feature set FS1. As features selection set was used with stacking using random tree, however, considerable improvement of accuracy were obtained.

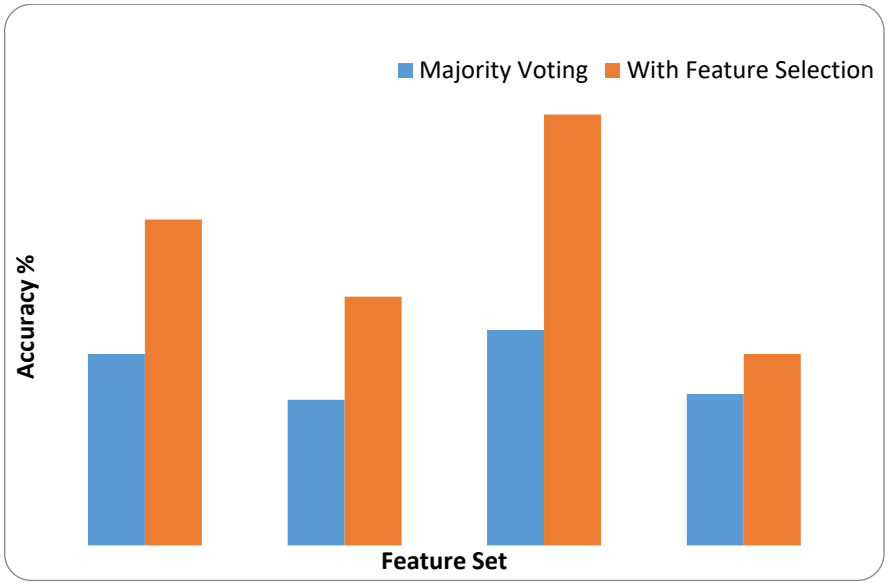

Figure 16. Improvement in precision of Majority Voting with RF, SVM, NB and BN using Feature set Selection.

Table 4summarizes the findings. When feature selection set FS4 was used with random tree, stacking of Naive Bayes, C4.5, Bayesian Network, SVM, and MLP with maximum increase of 4.63% was recorded.

Table 4. Improvement with Stacking method and Feature set Selection

Table 4. Improvement with Stacking method and Feature set Selection.			
Algorithms Stacked used with RT	Accuracy of stacking	Feature set Selection Accuracy	Feature set Selection
SVM, C4.5, Naïve Bayes	77.89	78.55	FS1
SVM, NaïveBayes,C4.5	77.89	78.55	FS2
SVM, C4.5, Naïve Bayes, MLP	77.56	78.22	FS1
SVM, C4.5, Naïve Bayes, MLP	77.56	77.89	FS3
SVM, C4.5, Naïve Bayes, MLP, Bayes Net	75.58	80.21	FS4
SVM, C4.5, Naïve Bayes, Bayes Net, MLP	75.58	76.24	FS2

CONCLUSION

The accuracy of cardiovascular disease prediction using various ensemble classifiers is investigated in this research. For training and testing, the Cleveland Heart dataset from the UCI Machine Learning Repository was used. The studies are carried out using a variety of ensemble approaches such as bagging, booting, stacking, and majority voting. When the bagging method is utilised, the precision improves by 5.92%. When the boosting method was used, accuracy increased by up to 6.54%. The accuracy of fragile classifiers improves by 6.96% when used with ensembled majority voting and by up to 7.13% when used with stacking. An analysis of the results shows that the voting mechanism nearly increases accuracy. When employed with a dataset, a feature selection process outperforms the previous results. As a result, the feature selection set aided in increasing the precision of ensemble methods. The highest accuracy was reached using the feature set FS4 via majority vote.

References

1. Latha CB, Jeeva SC. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked. 2019 Jan 1;16:100203.

2. Jothi KA, Subburam S, Umadevi V, Hemavathy K. Heart disease prediction system using machine learning. Materials Today: Proceedings. 2021 Feb 19.
3. Raza K. Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule. InU-Healthcare Monitoring Systems 2019 Jan 1 (pp. 179-196). Academic Press.
4. Sundari BA, Thanamani AS. An efficient feature selection technique using supervised fuzzy information theory. International Journal of Computer Applications. 2014 Jan 1;85(19).
5. Keerthana and Sundari. B.A., 2020. Heart Disease Data Pre-Processing Using Enhanced Data Mining Techniques. International Journal of Advanced Science and Technology 2020, 29(08), 6274-6282.
6. Yaman E, Subasi A. Comparison of bagging and boosting ensemble machine learning methods for automated EMG signal classification. BioMed research international. 2019 Oct 31;2019.
7. Radke RM, Frenzel T, Baumgartner H, Diller GP. Adult congenital heart disease and the COVID-19 pandemic. Heart. 2020 Sep 1;106(17):1302-9.
8. Saranya G, Pravin A. A comprehensive study on disease risk predictions in machine learning. International Journal of Electrical and Computer Engineering. 2020 Aug 1;10(4):4217.
9. Gupta A, Kumar R, Arora HS, Raman B. MIFH: A machine intelligence framework for heart disease diagnosis. IEEE access. 2019 Dec 27;8:14659-74.
10. Saqlain SM, Sher M, Shah FA, Khan I, Ashraf MU, Awais M, Ghani A. Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines. Knowledge and Information Systems. 2019 Jan;58(1):139-67.
11. Pan C, Poddar A, Mukherjee R, Ray AK. Impact of categorical and numerical features in ensemble machine learning frameworks for heart disease prediction. Biomedical Signal Processing and Control. 2022 Jul 1;76:103666.
12. Shah SM, Shah FA, Hussain SA, Batool S. Support vector machines-based heart disease diagnosis using feature subset, wrapping selection and extraction methods. Computers & Electrical Engineering. 2020 Jun 1;84:106628.
13. Haq AU, Li JP, Memon MH, Nazir S, Sun R. A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile Information Systems. 2018 Dec 2;2018.
14. Javid I, Alsaedi AK, Ghazali R. Enhanced accuracy of heart disease prediction using machine learning and recurrent neural networks ensemble majority voting method. International Journal of Advanced Computer Science and Applications. 2020;11(3).
15. Amin, M.S., Chiam, Y.K. and Varathan, K.D., 2019. Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics, 36, 2019. pp.82-93.